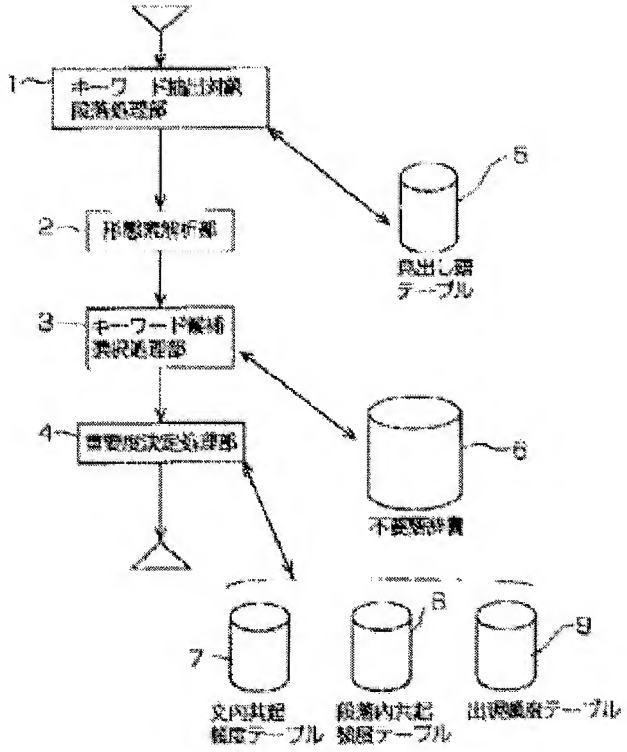
DEVICE AND METHOD FOR AUTOMATICALLY EXTRACTING KEYWORD

Publication number:	JP8202737
Publication date:	1996-08-09
Inventor:	HARA MASAMI
Applicant:	NTT DATA TSUSHIN KK
Classification:	
- international:	***G06F17/30;* G06F17/30;** (IPC1-7): G06F17/30
- European:	
Application number:	JP19950029949 19950126
Priority number(s):	JP19950029949 19950126

Abstract of **JP8202737**

PURPOSE: To automatically extract a high-quality keyword out of a text at high speed.
CONSTITUTION: This device is provided with a keyword extracting object paragraph end specifying processing part 1 for specifying any index word required as a keyword extracting object among index words registered on an index word table 5, morpheme analystic part 2 for dividing a sentence in the specified index word into words, keyword candidate selecting processing part 3 for collating the respective words with an unwanted word dictionary 6 and selecting only the required words as keyword candidates, and importance degree deciding processing part 4 for deciding the degrees of importance for the words defined as candidates based on their appearance frequency and including relation of character, sorting those words in order of descent from the highest degree of importance and defining the higher-order word as a keyword.



(19)日本国特許庁（ＪＰ）　(12) **公開特許公報**（Ａ）　(11)特許出願公開番号

特開平8－202737

(43)公開日　平成８年(1996)８月９日

(51)Int.Cl.[6]	識別記号	庁内整理番号	ＦＩ		技術表示箇所
Ｇ０６Ｆ　17/30					
		9194－5L	Ｇ０６Ｆ　15/ 401	３１０　Ｂ	
		9194－5L	15/ 40	３７０　Ａ	

審査請求　未請求　請求項の数６　ＦＤ　（全 11 頁）

(21)出願番号　特願平7－29949

(22)出願日　平成７年(1995)１月26日

(71)出願人　000102728
エヌ・ティ・ティ・データ通信株式会社
東京都江東区豊洲三丁目３番３号

(72)発明者　原　正巳
東京都江東区豊洲三丁目３番３号　エヌ・ティ・ティ・データ通信株式会社内

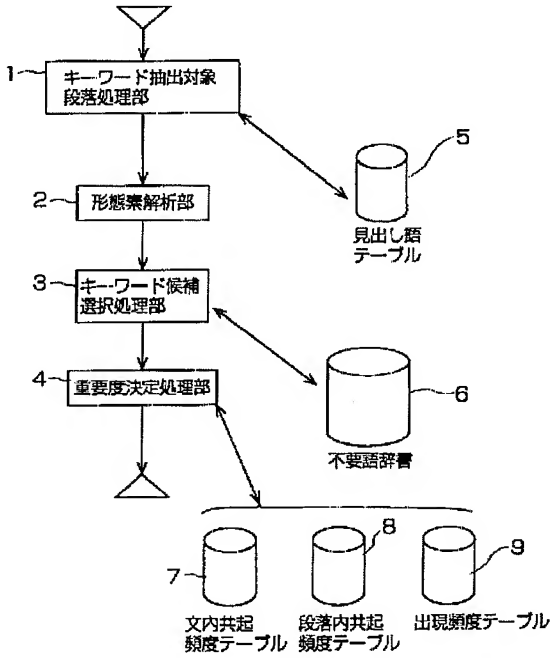
(74)代理人　弁理士　上村　輝之

(54)【発明の名称】　キーワード自動抽出装置およびキーワード自動抽出方法

(57)【要約】

【目的】　テキスト中から高品質のキーワードを高速に自動抽出する。

【構成】　見出し語テーブル５に登録されている見出し語の内キーワード抽出対象として必要なものを特定するキーワード抽出対象段落特定処理部１と、特定した見出し語内の文を単語に分割する形態素解析部２と、各単語に対して不要語辞書６との照合を行い、必要な単語のみキーワード候補として選択するキーワード候補選択処理部３と、候補とされた単語に対してその出現頻度と文字の包含関係に基づき重要度を決定し、重要度の高い順にソートして上位の単語をキーワードとする重要度決定処理部４とを備えた。



1

【特許請求の範囲】

【請求項１】　予め定めた見出し語を登録した見出し語テーブルと、

テキストのデータを入力し、前記テキスト中の段落の内から、前記見出し語テーブルに登録されている見出し語のいずれかを含んだ段落を、キーワード抽出対象段落として特定するキーワード抽出対照段落特定処理部とを備え前記特定されたキーワード抽出対照段落からキーワード抽出を行うことを特徴とするキーワード自動抽出装置。

【請求項２】　テキストのデータを入力して、このテキストを単語に分割する形態素解析部と、

予め定めた不要語を登録した不要語辞書と、

前記形態素解析部で得られた各単語に対して不要語辞書との照合を行い、必要な単語のみキーワード候補として選択するキーワード候補選択処理部とを備え、

前記選択されたキーワード候補の中からキーワード抽出を行うようにしたことを特徴とするキーワード自動抽出装置。

【請求項３】　テキストのデータを入力して、このテキストの中からキーワード候補を選択する選択処理部と、

前記選択された各キーワード候補について、前記テキスト内での出現頻度に関する統計量を計算する頻度計算部と、

前記計算された各キーワード候補の統計量を記録した頻度テーブルと、

前記頻度テーブルに記録された各キーワード候補の統計量から、各キーワード候補に対して重要度を決定し、重要度に基づいて前記キーワード候補中からキーワードを抽出する重要度決定処理部とを備えたことを特徴とするキーワード自動抽出装置。

【請求項４】　請求項３記載の装置において、

前記統計量として、前記テキスト内での各キーワード候補それ自体の出現頻度と、前記テキストを区分した所定範囲で各キーワード候補と他のキーワード候補とが共に出現する頻度たる共起頻度と、キーワード候補同士の含有関係を利用した最長語への重要度補正とが用いられることを特徴とするキーワード自動抽出装置。

【請求項５】　予め定めた見出し語を登録した見出し語テーブルと、

テキストのデータを入力し、前記テキスト中の段落の内から、前記見出し語テーブルに登録されている見出し語のいずれかを含んだ段落を、キーワード抽出対象段落として特定するキーワード抽出対照段落特定処理部と前記キーワード抽出対照段落を単語に分割する形態素解析部と、

予め定めた不要語を登録した不要語辞書と、

前記形態素解析部で得られた各単語に対して不要語辞書との照合を行い、必要な単語のみキーワード候補として選択するキーワード候補選択処理部と前記選択された各

2

キーワード候補について、前記テキスト内での出現頻度に関する統計量を計算する頻度計算部と、

前記計算された各キーワード候補の統計量を記録した頻度テーブルと、

前記頻度テーブルに記録された各キーワード候補の統計量から、各キーワード候補に対して重要度を決定し、重要度に基づいて前記キーワード候補中からキーワードを抽出する重要度決定処理部とを備えたことを特徴とするキーワード自動抽出装置。

【請求項６】　テキスト中の段落の中から、見出し語テーブルに登録されている見出し語を含む段落をキーワード抽出対象段落として特定する第１の工程と、

この第１の工程で特定したキーワード抽出段落を単語に分割する第２の工程と、

この第２の工程で得られた各単語に対して不要語辞書との照合を行い、キーワード候補を選択する第３の工程と、

この第３の工程で候補とされた単語に対して重要度を決定し、重要度の高い単語をキーワードとする第４の工程と、

を有することを特徴とするキーワード自動抽出方法。

【発明の詳細な説明】

【０００１】

【産業上の利用分野】本発明は、テキスト中のキーワードを自動的に抽出するキーワード自動抽出装置およびキーワード自動抽出方法に関する。

【０００２】

【従来の技術】テキスト中のキーワードの抽出は、従来、人間がテキストを熟読し、内容を熟知した上で手作業で行っていた。

【０００３】しかし、テキストの電子化が進み、膨大な数量でかつ長大なテキストデータを扱う必要が生じてきた現在、キーワードの作成を人手で行うことは事実上不可能となっている。

【０００４】そこで、このような電子化されたテキストに対して、コンピュータによりキーワードを自動的に抽出する方法が検討されてきている。

【０００５】その方法として、自然言語処理技術、特に文の意味をコンピュータ上で解釈する意味理解技術を用いて文中の各語の重要性を決定する方法や、出現頻度、語長、文字種などテキストの表層情報を利用することにより重要性を決定する方法が考えられている。

【０００６】

【発明が解決しようとする課題】しかしながら、膨大な数量でかつ長大なテキストに対して、意味解析や文脈解析などの自然言語処理を実施することは現状では困難である。従って、上述の意味理解技術を用いたキーワード自動抽出を高い精度で達成することは困難であり、また、仮にキーワード自動抽出を実現しても多大な実行時間を要するという問題があった。

【０００７】一方、表層情報を利用して重要性を決定するキーワード自動抽出方式においては、高速処理は実現できるものの、語の意味や語同士の関連性を考慮していないため、実際には余り重要でない語がキーワードとして抽出されやすいという問題があった。また、必要な語句がキーワードとして抽出されない場合が生じるという不具合もあった。

【０００８】本発明はこのような背景に基づいてなされたものであり、その目的は、テキスト中から高品質なキーワードを高速に自動抽出することにある。

【０００９】

【課題を解決するための手段】上記の目的を達成するために、本発明の第１の側面に従うキーワード自動抽出装置は、予め定めた見出し語を登録した見出し語テーブルと、テキストのデータを入力し、テキスト中の段落の内から、見出し語テーブルに登録されている見出し語のいずれかを含んだ段落を、キーワード抽出対象段落として特定するキーワード抽出対照段落特定処理部とを備え、特定されたキーワード抽出対照段落からキーワード抽出を行うことを特徴とする。

【００１０】本発明の第２の側面に従うキーワード自動抽出装置は、テキストのデータを入力して、このテキストを単語に分割する形態素解析部と、予め定めた不要語を登録した不要語辞書と、形態素解析部で得られた各単語に対して不要語辞書との照合を行い、必要な単語のみキーワード候補として選択するキーワード候補選択処理部とを備え、選択されたキーワード候補の中からキーワード抽出を行うようにしたことを特徴とする。

【００１１】本発明の第３の側面に従うキーワード自動抽出装置は、テキストのデータを入力して、このテキストの中からキーワード候補を選択する選択処理部と、選択された各キーワード候補について、テキスト内での出現頻度に関する統計量を計算する頻度計算部と、計算された各キーワード候補の統計量を記録した頻度テーブルと、頻度テーブルに記録された各キーワード候補の統計量から、各キーワード候補に対して重要度を決定し、重要度に基づいてキーワード候補中からキーワードを抽出する重要度決定処理部とを備えたことを特徴とする。

【００１２】本発明の第４の側面に従うキーワード自動抽出装置は、予め定めた見出し語を登録した見出し語テーブルと、テキストのデータを入力し、テキスト中の段落の内から、見出し語テーブルに登録されている見出し語のいずれかを含んだ段落を、キーワード抽出対象段落として特定するキーワード抽出対照段落特定処理部と、キーワード抽出対照段落を単語に分割する形態素解析部と、予め定めた不要語を登録した不要語辞書と、形態素解析部で得られた各単語に対して不要語辞書との照合を行い、必要な単語のみキーワード候補として選択するキーワード候補選択処理部と、選択された各キーワード候補について、テキスト内での出現頻度に関する統計量を計算する頻度計算部と、計算された各キーワード候補の統計量を記録した頻度テーブルと、頻度テーブルに記録された各キーワード候補の統計量から、各キーワード候補に対して重要度を決定し、重要度に基づいてキーワード候補中からキーワードを抽出する重要度決定処理部とを備えたことを特徴とする。

【００１３】本発明の第５の側面に従うキーワード自動抽出方法は、テキスト中の段落の中から、見出し語テーブルに登録されている見出し語を含む段落をキーワード抽出対象段落として特定する第１の工程と、この第１の工程で特定したキーワード抽出段落を単語に分割する第２の工程と、この第２の工程で得られた各単語に対して不要語辞書との照合を行い、キーワード候補を選択する第３の工程と、この第３の工程で候補とされた単語に対して重要度を決定し、重要度の高い単語をキーワードとする第４の工程とを有することを特徴とする。

【００１４】

【作用】本発明の第１の側面に係る装置は、テキストに含まれる段落の内、見出し語テーブルに予め登録されている見出し語を備えた段落だけを、キーワード抽出対象段落として特定し、この特定したキーワード抽出対照段落からキーワード抽出を行う。そのため、キーワードが含まれている可能性の低い段落からキーワード抽出する無駄が省かれる。

【００１５】また本発明の第２の側面に係る装置は、形態素解析部で得られたテキスト中の各単語に対して不要語辞書との照合を行い、必要な単語のみをキーワード候補として選択し、キーワード候補とされた単語の中からキーワード抽出を行う。そのため、キーワードとなり得ない不要な単語をも含んだ膨大なデータに対してキーワード抽出処理を行う無駄が省かれる。

【００１６】また本発明の第３の側面に係る装置は、テキストの中からキーワード候補を選択し、キーワード候補とされた単語に対して、出現頻度に基づく重要度を決定し、重要度の高い単語を優先的にキーワードとする。そのため、キーワードである確率の低い単語が除外され、キーワード抽出の精度が高まる。

【００１７】ここで、重要度は、キーワード候補の出現頻度だけでなく、他のキーワード候補との文字の含有関係をも考慮して決定することが望ましい。その場合、統計量としては、例えば、テキスト内での各キーワード候補それ自体の出現頻度と、テキストを区分した所定範囲で各キーワード候補と他のキーワード候補とが共に出現する頻度である共起頻度と、更に、キーワード候補同士の含有関係を利用した最長語への重要度補正とを用いることができる。このように出現頻度と文字の含有関係とに基づき重要度を決定することにより、より一層の精度向上が期待できる。一般に、キーワードに適した重要単語は、出現頻度が高い傾向があり、さらに、その重要単語の近傍に現れる語は、重要単語と密接に関連してテキ

ストの主題を表現する傾向があるため、キーワードになり易いからである。

【００１８】また、この場合、キーワードの部分一致による重要度補正では最長の単語を優先することが好ましい。一般に、長い語句ほどより内容が限定されることと、同一テキストにおいて部分的に一致する単語は、最長の単語の内容をより抽象的に述べていることが多いからである。

【００１９】また本発明の第４の側面に係る装置又は第５の側面に係る方法によれば、見出し語テーブルに登録されている見出し語を備える段落がキーワード抽出対象段落として特定され、特定されたキーワード抽出対照段落が単語に分割される。次に、各単語に対して不要語辞書との照合が行われ、必要な単語のみがキーワード候補として選択され、次いで、キーワード候補とされた単語に対して重要度が決定され、重要度の高い単語がキーワードとして選択される。このため、処理の早い段階で不要なデータが除外されて処理負担が減るために、処理速度が向上すると共に、キーワードである可能性の高いデータだけを抽出するフィルタリングが異なる観点から複数段階にわたって行われるため、キーワード抽出の精度が向上する。

【００２０】

【実施例】以下、本発明の一実施例を添付図面に基づいて詳細に説明する。

【００２１】図１は本実施例に係るキーワード自動抽出装置の機能ブロック図である。

【００２２】この図において、１はキーワード抽出対象段落特定処理部（以下、単に段落特定処理部と称する）である。この段落特定処理部１は見出し語テーブル５と信号の授受を行うようになっている。段落特定処理部１の機能については後述する。

【００２３】２は形態素解析部である。この形態素解析部２では文を単語に分割する。

【００２４】３はキーワード候補選択処理部（以下、単に候補選択処理部と称する）である。この候補選択処理部３は不要語辞書６と信号の授受を行うようになっている。候補選択処理部３の機能については後述する。

【００２５】４は重要度決定処理部である。この重要度決定処理部４は文内共起頻度テーブル７、段落内共起頻度テーブル８、出現頻度テーブル９のそれぞれと信号の授受を行うようになっている。重要度決定処理部４の機能については後述する。

【００２６】図２は段落特定処理部１における制御動作のフローチャートである。

【００２７】動作を、定型フォーマットのテキストの例として特許明細書を用いて説明する。

【００２８】まず、特許明細書のデータが入力されると、段落特定処理部１が起動される。段落特定処理部１ではテキストから１行を読み込み（Ｓ１）、見出し語テーブル５を参照して、見出し語を含むかどうかを調べる（Ｓ２）。見出し語テーブル５を参照した結果、見出し語が存在しなければ（Ｓ２でＮ）、直前の行と同様の処理を行う（Ｓ３）。但し１行目については、見出し語が存在しない場合スキップする。

【００２９】図３は特許明細書における見出し語を示す説明図である。

【００３０】「発明の名称」、「構成」、「産業上の利用分野」等の見出し語には、要、不要のマークが「１」、「０」として付されている。キーワードが含まれている可能性がある見出し語、即ち、キーワード自動抽出に必要な見出し語は「１」が付されており、そうでない見出し語は「０」が付されている。例えば、見出し語「発明の名称」はキーワード自動抽出に必要であり、「産業上の利用分野」は必要でない。

【００３１】再び図２のフローチャートに戻り、見出し語が存在した場合（Ｓ２でＹ）、不要な見出し語でなければ（Ｓ４でＹ）、キーワード抽出対象として採用する（Ｓ６）。一方、不要な見出し語であれば（Ｓ４でＮ）、スキップする（Ｓ５）。採用された行はその見出し語に属する文として追加される。

【００３２】図４は必要な見出し語とそれに属する文を示す説明図である。

【００３３】例えば、必要な見出し語として挙げられている「発明の名称」に属する文は「キーワード自動抽出方式」であることが示されている。

【００３４】再び図２のフローチャートに戻り、不要な見出し語が存在した後は、次に必要な見出し語が現れるまで（Ｓ４でＹ）、Ｓ５、Ｓ１、Ｓ２のルーチンが繰り返される。

【００３５】以上の処理をテキストが終了するまで（Ｓ７でＹ）行う。

【００３６】形態素解析部２では、段落特定処理部１で得られた見出し語内の文を単語に分割する。

【００３７】図５は見出し語内の文とその文の単語を示す説明図である。

【００３８】「各確率的予測関数・・・計算する」という文が、“各”、“確率的予測関数”、・・・“計算”、“する”等の単語に分割される。

【００３９】図６は候補選択処理部３における制御動作のフローチャートである。

【００４０】候補選択処理部３では、形態素解析部２により単語切りされた各語を取り込んで（Ｓ１１）、この語について不要語辞書６を照合し（Ｓ１２）、不要語辞書に登録されている語は削除し（Ｓ１３）、それ以外はキーワード候補とする（Ｓ１４）。形態素解析部２により単語切りされた全単語について上述の処理が終了した時点で（Ｓ１５でＹ）、このフローは終了する。

【００４１】図７は候補選択処理部３の出力例を示す説明図である。

【００４２】例えば、見出し語「特許請求の範囲」の段落の文中、“定型フォーマット”、“テキスト”等がキーワード候補として挙げられている。

【００４３】図８は重要度決定処理部４における制御動作のフローチャートである。

【００４４】重要度決定処理部４では、候補選択処理部３により候補とされた語について、まず同一文内での共起頻度を求め、文内共起頻度テーブル７に登録する（Ｓ２１、Ｓ２２）。次に、同一見出し語内での共起頻度を求め、段落内共起頻度テーブル８に登録する（Ｓ２３）。さらに、テキスト全体における語単独の出現頻度を求め、出現頻度テーブル９に登録する（Ｓ２４）。

【００４５】以上の処理を処理対象段落がなくなるまで（Ｓ２５でN）実行する。

【００４６】図９は共起頻度テーブルの一例を示す説明図である。

【００４７】この図において、「確率分布」は「解析システム」とは同時に出現はせず、また「微分方程式」とは９回同時に出現することが示されている。さらに合計により、「確率分布」が他の語と共起して出現する回数は２０回であることが示されている。

【００４８】再び図８のフローチャートに戻り、処理対象段落を全て処理した後、作成された文内共起頻度テーブル７と段落内共起頻度テーブル８で求められた共起頻度および出現頻度テーブル９で求められたテキスト全体の出現頻度の合計を基にして、重要度Ｉが決定される（Ｓ２６）。

【００４９】図１０は重要度の算出の仕方を示す説明図である。

【００５０】重要度Ｉは、

Ｉ＝α・（共起頻度テーブル７における各単語の合計値）＋β・（共起頻度テーブル８における各単語の合計値）＋（出現頻度テーブル９の合計値）

で表される。α，βは定数である。

【００５１】ここで、α＝３、β＝２とした場合、例えば“確率分布”の重要度Ｉは、

Ｉ＝３×８＋２×３２＋２３＝１１１

ということになる。また同様に“情報管理”の重要度Ｉは８９ということになる。

【００５２】このようにして、図８のステップＳ２６において、各単語の重要度は決定される。次にキーワード候補語の含有関係を調査し、語長の長いキーワード候補語に含まれる語が、同様にキーワード候補語に含まれる場合、重要度の補正を行う（Ｓ２７）。

【００５３】図１１は補正された重要度の算出の仕方を示す説明図である。

【００５４】補正重要度Ｉ＊は、

Ｉ＊＝（語長の長いキーワード候補語の重要度Ｉ）＋γ・（長い候補語に含まれる候補語の重要度Ｉ）

によって求められる。γは定数である。

【００５５】例えば、γ＝１とした場合、“確率分布”の場合、重要度は前述したように１１１であるが、“確率”の重要度は４２であるので、“確率分布”の補正重要度は１５３（＝１１１＋４２）ということになる。

【００５６】再び図８のフローチャートに戻り、このようにして補正された重要度の高い順に単語をソートし、上位の語をキーワードとする（Ｓ２８）。

【００５７】上述した一連の処理を実行することにより、キーワードの自動抽出を高速に、かつ効率的に行うことができる。

【００５８】本実施例は、段落特定処理部１、形態素解析部２、候補選択処理部３、重要度決定処理部４の各処理過程を経て、キーワード自動抽出を行うようにしているが、この内の一つの処理だけを採用しても、従来例に比べて高速に処理することができる。

【００５９】例えば、段落特定処理部１を用い、予め重要な語句を入りやすい段落の見出し語を調査しておくことで、不要な段落に関する処理を回避し、高速にキーワードを抽出することができる。

【００６０】また、候補選択処理部３を用い、予め不要な単語は削除しておくだけでもキーワード自動抽出の高速化を図ることができる。

【００６１】さらに、重要度決定処理部４で、表層情報である出現頻度や共起出現頻度、語の含有関係を総合的に判断することにより従来のように、複雑かつ長時間にわたりテキストの意味や文脈を解析することを回避し、かつ語同士の関連を考慮したキーワード抽出が可能となる。

【００６２】なお、本実施例ではテキストとして特許明細書を挙げて説明したが、他の定型フォーマットのテキストにも適用できることは言うまでもない。

【００６３】

【発明の効果】本発明によれば、高速にキーワードの自動抽出を行うことができる。

【図面の簡単な説明】

【図１】本発明の一実施例に係るキーワード自動抽出装置の機能ブロック図である。

【図２】キーワード抽出対象段落特定処理部における制御動作のフローチャートである。

【図３】特許明細書における見出し語を示す説明図である。

【図４】必要な見出し語とそれに属する文を示す説明図である。

【図５】見出し語内の文とその文の単語を示す説明図である。

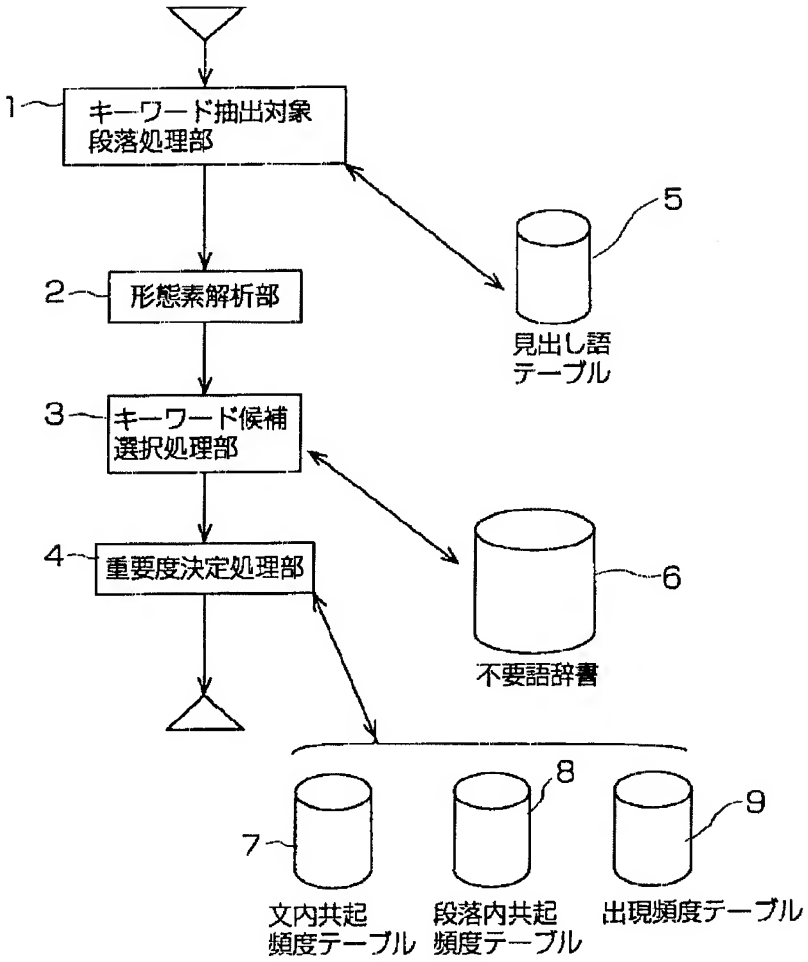
【図６】キーワード候補選択処理部における制御動作のフローチャートである。

【図７】キーワード候補選択処理部の出力例を示す説明図である。

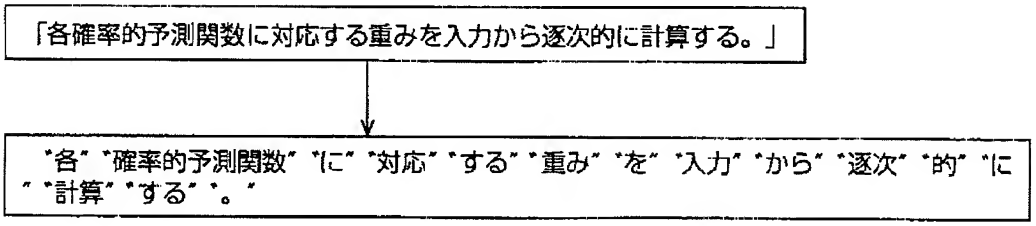
【図８】重要度決定処理部における制御動作のフローチ

ャートである。
【図９】共起頻度テーブルの一例を示す説明図である。
【図１０】重要度の算出の仕方を示す説明図である。
【図１１】補正重要度の算出の仕方を示す説明図である。
【符号の説明】
1　キーワード抽出対象段落特定処理部
2　形態素解析部
＊3　キーワード候補選択処理部
4　重要度決定処理部
5　見出し語テーブル
6　不要語辞書
7　文内共起頻度テーブル
8　段落内共起頻度テーブル
9　出現頻度テーブル
＊

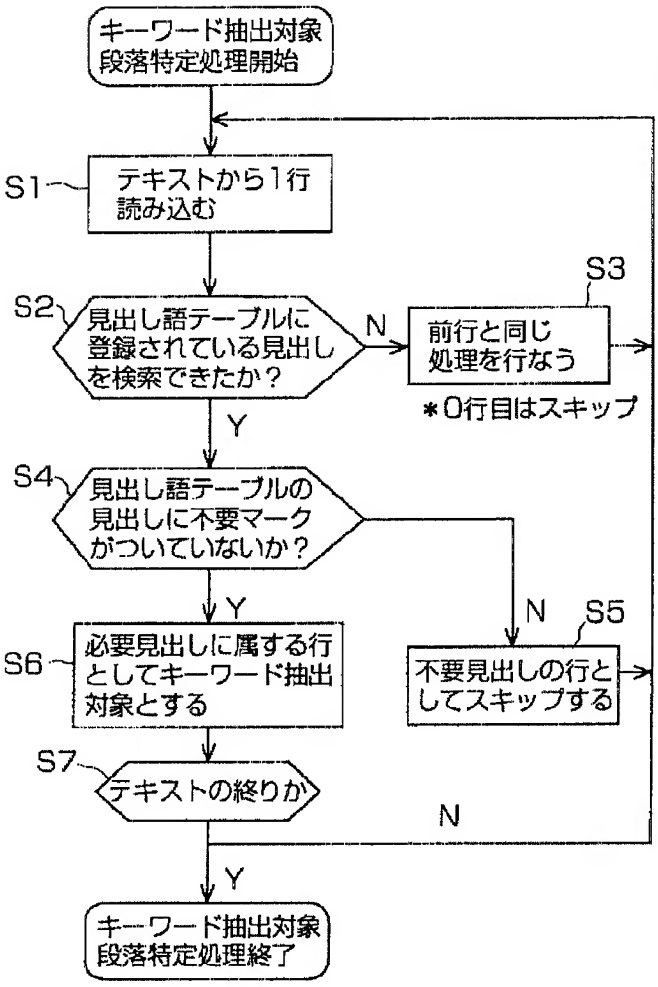
【図１】



【図５】



【図2】



【図9】

	確率分布	解析システム	微分方程式	経済分析	…
確率分布	×	0	9	1	：
解析システム	0	×	5	3	：
微分方程式	9	5	×	7	：
経済分析	1	3	7	×	：
：	：	：	：	：	：
合計	20	10	28	16	…

【図4】

【発明の名称】
　　"キーワード自動抽出方式"
【特許請求の範囲】
　　"定型フォーマットを持つテキスト中から重要パラグラフを特定する重要パラグラフ特定手段と…"
【課題を解決するための手段】
　　"上記課題を解決するために、…"
【発明の効果】
　　"以上説明したように、本発明によれば、人手による作成労力の減少…。"
　　"本発明では、特許明細書を例に説明したが、他の…。"

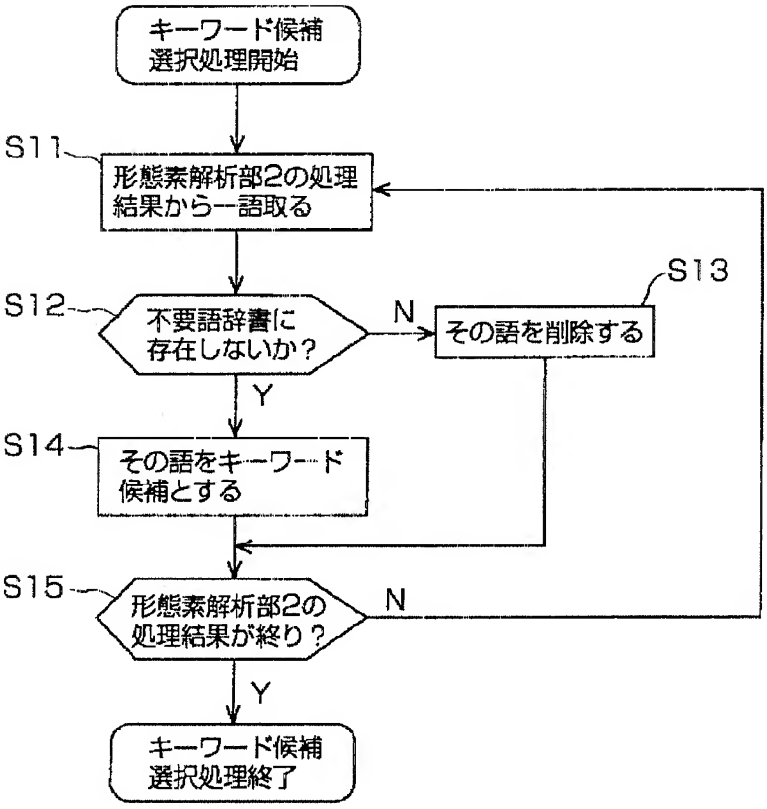
【図３】

見出し語	要／不要
【発明の名称】	1
【目的】	1
【構成】	1
【効果】	1
【特許請求の範囲】	1
【産業上の利用分野】	0
【技術分野】	0
【従来の技術】	0
【従来技術】	0
【発明が解決しようとする課題】	1
【課題を解決するための手段】	1
【作用】	0
【発明の目的】	0
【発明の概要】	0
【実施例】	0
【発明の効果】	1
【図面の簡単な説明】	0
【符号の説明】	0
:	:

【図７】

【発明の名称】
　　"キーワード自動抽出方式"
【特許請求の範囲】
　　"定型フォーマット" "テキスト" "重要パラグラフ" "重要パラグラフ特定手段"
【課題を解決するための手段】
　　"自然言語処理" "形態素解析" "定型フォーマット"
【発明の効果】
　　"作成労力" "減少" "意味理解" "言語理解"

【図６】



【図１１】

	確率分布	情報管理	経済分析	確率	…
重要度	111	89	116	42	…

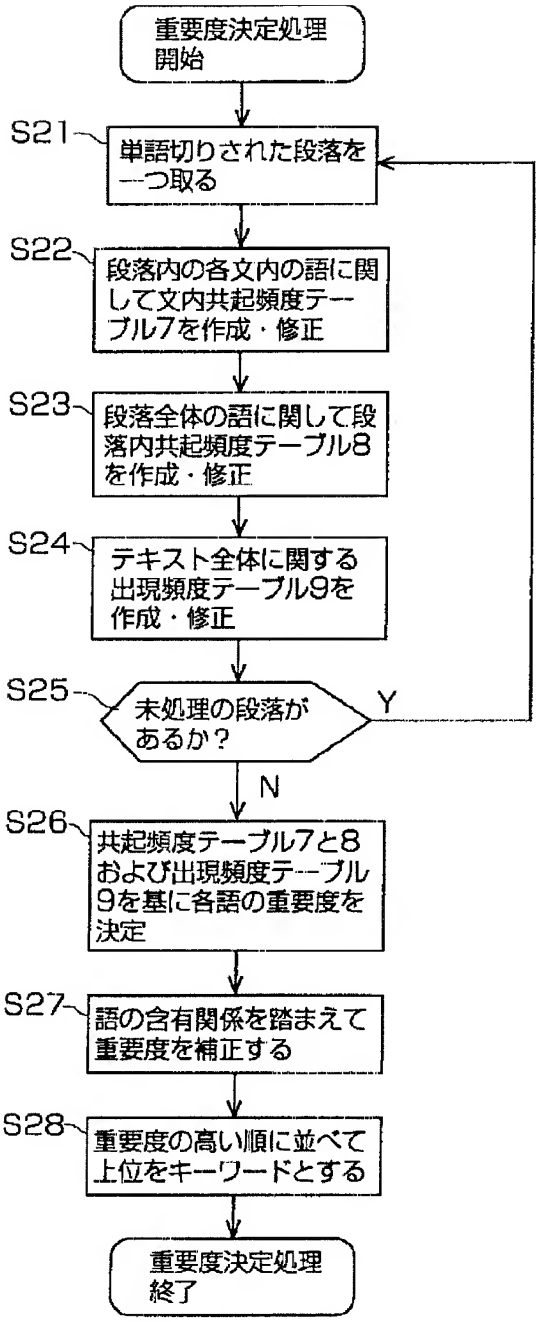
⇓

	確率分布	微分方程式	経済分析	…
重要度	153	39	126	…

重要度I^*＝(長い語句Wの重要度Iw)＋γ・(Wに含まれる語句wの重要度Iw)　(γ：定数)

例はγ＝１の場合

【図８】


重要度決定処理開始
S21
単語切りされた段落を一つ取る
S22
段落内の各文内の語に関して文内共起頻度テーブル7を作成・修正
S23
段落全体の語に関して段落内共起頻度テーブル8を作成・修正
S24
テキスト全体に関する出現頻度テーブル9を作成・修正
S25
未処理の段落があるか？
Y
N
S26
共起頻度テーブル7と8および出現頻度テーブル9を基に各語の重要度を決定
S27
語の含有関係を踏まえて重要度を補正する
S28
重要度の高い順に並べて上位をキーワードとする
重要度決定処理終了

【図１０】

共起頻度テーブル7（文内での共起数）

	確率分布	情報管理	経済分析	確率	…
確率分布	×	5	1	2	:
情報管理	5	×	7	0	:
経済分析	1	7	×	1	:
確率	2	0	1	×	:
:	:	:	:	:	:
合計	8	14	6	3	…

共起頻度テーブル8（段落内での共起数）

	確率分布	情報管理	経済分析	確率	…
確率分布	×	9	15	4	:
情報管理	9	×	7	1	:
経済分析	15	7	×	2	:
確率	4	1	2	×	:
:	:	:	:	:	:
合計	32	19	28	9	…

α・（各単語の共起数の合計）

β・（各単語の共起数の合計）

出現頻度テーブル9（テキスト全体での出現頻度）

	確率分布	情報管理	経済分析	確率	…
出現頻度	23	9	42	3	…

⇓

	確率分布	情報管理	経済分析	確率	…
重要度	111	89	116	42	…

重要度I＝α・（各単語の共起頻度テーブル7における合計）
＋β・（各単語の共起頻度テーブル8における合計）
＋（出現頻度テーブル9による頻度）　　（α, β：定数）

例はα＝3, β＝2の場合